Universitat de les Illes Balears

Modelación y predicción del gasto de turistas en España enfocado desde el análisis de datos

MSC Candidate

Laura Piña Miranda

A MSc thesis submitted to the University of Balearic Islands in accordance with the requirements for the degree of **Màster Universitari en Anàlisi de Dades Massives en Economia i Empresa (MADM)**

Author _____

Certified by _____

Jaume Rosselló Nadal
UIB Master's Thesis Supervisor

Certified by _____

Antonio Bibiloni
Master's Thesis Co Supervisor

24 Septiembre 2018

Quiero agradecer con mucho cariño el apoyo incondicional de mi esposo y mis padres, así como mi familia que también de una manera indirecta influyeron en mi constancia para terminar el Máster. A mis tutores porque estuvieron pendientes de mi trabajo y me dieron sus valoraciones positivas para concluirlo. A mis amigos que también confiaron en mi perseverancia y entusiasmo.

Modelación y predicción del gasto de turistas en España enfocado desde el análisis de datos

Laura Piña Miranda
Tutor: Jaume Rosselló Nadal & Antonio Bibiloni

Treball de fi de Màster Universitari en Analisi de Dades Massives en Economia i Empresa (MADM)

Universitat de les Illes Balears
07122 Palma de Mallorca
laura_pia2002@yahoo.com

Resumen

Se crearon modelos predictivos para el gasto total de los turistas en España utilizando los datos de la encuesta EGATUR. Se utilizaron los algoritmos de Regresión Lineal (LR), Random Forest (RF), Máquina de vectores de soporte (SVM) y Aprendizaje Profundo (DL). Se buscó para cada algoritmo el conjunto de sus hiperparámetros óptimos. Se estimó la capacidad predictiva de cada método utilizando los métodos de validación cruzada y set de validación. Los algoritmos con mayor exactitud en sus predicciones son DL y SVM, siendo este último mucho más lento. Los residuos obtenidos para cada par de métodos correlacionan entre sí, por lo que no es posible obtener mejores predicciones promediando los modelos obtenidos. Se analizó la importancia de las variables, siendo el número de pernoctaciones la más importante seguido del país de origen.

Abstract

Predictive models for the total expenditure of tourist in Spain were made by using the data of the EGATUR survey. The Linear Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Deep Learning (DL) algorithms were utilized. For each algorithm a search for the optimal hyperparameters was performed. The predictive capabilities of each method was assessed by means of the cross validation and validation set methodologies. The algorithms with best predictive capabilities were DL and SVM, being the latter a lot slower. The residuals for each pair of methods correlates, then averaging the models doesn´t yield better predictions. The importance of each variable was analyzed, being the number of nights to stay the most relevant followed by the country of origin of the tourist.

Palabras clave: Turismo en España, gasto total, predicción, bidgata,

1. Introducción

El análisis de datos masivos, también denominado como “Big Data” se basa en almacenar y posteriormente cuantificar y procesar grandes cantidades de información con el objetivo de obtener tendencias y regularidades. Estas metodologías se han convertido en los últimos años en una de las herramientas más utilizadas tanto para la investigación científica básica como de carácter comercial, siendo empleadas en sectores tan disímiles como la física[1-3], medicina[4,5], economía[6-8] y aviación[9,10] entre otros[11].

Entre las posibles aplicaciones del análisis de datos en la investigación básica, se encuentra la obtención de nuevos modelos matemáticos que relacionan diferentes parámetros a partir de un gran conjunto de resultados experimentales. Estos modelos, en los casos en que sean relativamente sencillos pueden ser interpretados con las teorías existentes[12] (de forma que las validen o las refuten), así como pueden dar surgimiento a nuevas hipótesis teóricas[13]. En el campo económico, el análisis de datos permite entre otras cosas, analizar los factores de una actividad económica que mayor impactan sus resultados[14]. De esta forma es posible obtener predicciones con mayor grado de confianza del efecto de las diferentes decisiones comerciales posibles permitiendo optimizar el camino a seguir en busca del éxito comercial. Esto muchas veces conlleva, para las empresas que aplican el análisis de datos, una ventaja competitiva respecto a las que no lo hacen.

La creciente relevancia de este campo en la actualidad está dada por la mayor facilidad de la obtención y almacenamiento de datos así como por el aumento del poder de procesamiento de las

computadoras. El primer factor responde al cada vez más fácil acceso a internet, siendo esta la llamada era digital. Esto abarata y agiliza el proceso de reunir grandes volúmenes de datos, ya sea por medio de encuestas online, el uso de las redes sociales, datos almacenados por buscadores y más recientemente datos provistos por los electrodomésticos inteligentes conectados a la web. De igual forma, el constante incremento en los últimos años del poder de cálculo de las computadoras ha permitido la aplicación de algoritmos cada vez más complejos y sofisticados para encontrar patrones subyacentes en grandes conjuntos de observaciones. Estos dos factores han conllevado una mayor valoración de grandes volúmenes de datos y del hombre que los analiza, el "analista de datos", habiéndose declarado esta ocupación como el mejor trabajo del año en el 2018 en los Estados Unidos[15].

El turismo ha devenido como uno de las actividades impulsoras de la economía que más interés ha despertado en las últimas décadas. Es la cuarta mayor industria de exportación representando el 6% de las exportaciones de mercancía y servicios del mundo[16]. Producto de esto, han sido varios los estudios realizados de su impacto en el crecimiento económico de diferentes países[17-20]. Esta actividad económica genera valor añadido al patrimonio tangible e intangible gracias a los bienes y servicios que se ofrecen, los cuales comprenden un amplio espectro de industrias y actividades cuyo único factor en común es el turista como receptor. De esta manera, el turismo permite ingresar al país divisas extranjeras, fomenta la inversión en infraestructura y permite incorporar un gran volumen de recursos humanos, representando esta actividad una de las principales fuentes de empleo en la actualidad[21,22].

Existen varios factores que han impulsado el auge de esta actividad económica, siendo uno de los fundamentales el impacto de la globalización, pues cada vez es más sencillo transportarse de un país a otro. El caso de la Unión Europea es extremo, siendo uno de los mercados turísticos con más rápida expansión en el mundo. Por un lado esto se debe al creciente interés de los ciudadanos de los EEUU y Japón de visitar estados miembros. Por otro lado, producto de los acuerdos de libre movilidad, cada vez aumenta el turismo europeo en la zona Euro, trayendo una redistribución de las riquezas entre los estados miembros, muy provechosa para países con una menor industrialización como España, Portugal y Grecia. Es cada vez más común para los europeos hacer turismo por periodos cortos pero varias veces al año[23].

Dentro de los países de la unión, España destaca como destino turístico. El número de visitantes que arriba a esta nación ha estado en constante crecimiento, llegando a 82 millones de visitantes en el 2017, esto representó un ingreso de 87 mil millones de euros[24], equivalente al 16% de su PIB[25]. De igual forma algunas localidades como Mallorca, Menorca, Ibiza, donde la industrialización es relativamente baja generan, gracias al turismo de sol y playas importantes ingresos.

Dada la importancia de esta actividad económica para este país se hace necesario sacar el mayor provecho de la infraestructura existente así como saber qué características u ofertas potenciar. Para esto resulta muy atractivo utilizar herramientas del análisis de datos. Mediante éstas sería posible hallar regularidades complejas no evidentes así como se podrían crear herramientas predictivas con el fin de buscar los factores que respondan a preguntas tales como: ¿qué buscan los turistas? ¿qué turistas gastan más dinero? entre otras, que permitan tomar decisiones para hacer más atractivas las ofertas al tipo de turista óptimo, el que más dinero gasta.

En este trabajo se tiene como objetivo crear modelos predictivos del gasto de un turista en España. Para esto se utilizaron los datos de la encuesta EGATUR[26] comprendidos entre octubre de 2015 y agosto de 2017. Esta data reúne diferentes datos de turistas tales como tipo de alojamiento, número de días, entre otros factores cuya relación con el gasto total del turista no es evidente. Utilizaremos los métodos de regresión lineal (LR), k nearest neighbors (KNN), random forest (RF), deep learning (DL) y support vector machine (SVM) en su variante de regresión.

Regresión Lineal

La regresión lineal es uno de los algoritmos más simples de aprendizaje supervisado. Su descubrimiento se debe al matemático Fischer[27] en 1936 y continúa siendo en la actualidad uno de los métodos de ajuste más utilizados.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (0.1)$$

La expresión fundamental de la regresión simple se muestra en la ecuación (0.1). El objetivo es encontrar valores de β_0 (intercepto) y β_1 (pendiente) de forma tal que la ecuación de la línea recta obtenida ajusta la data lo mejor posible, o lo que es equivalente, se minimizan los errores o residuos $\{\varepsilon_i\}$. Existen varias vías para lograr este objetivo, no obstante la más utilizada y casi convertida en estándar es usar el criterio de mínimos cuadrados, es decir minimizar la suma de los cuadrados de los residuos (ecuación (0.2)).

$$RSS = \varepsilon_1^2 + \varepsilon_2^2 + ... + \varepsilon_n^2 \qquad (0.2)$$

Sustituyendo (0.1) en (0.2) se obtiene luego de operaciones de cálculo sencillas, las dos expresiones que dan los valores de β_0 y β_1 de forma directa (ecuación (0.3)), siendo por tanto este método uno de los más simples de evaluar.

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (0.3)$$
$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

K vecinos más cercanos

La técnica k vecinos más cercanos (KNN por sus siglas en inglés) es un algoritmo muy sencillo utilizado en problemas de clasificación y regresión inventado en la década de 1950 por Fix y Hodges[28] y continúa siendo en la actualidad uno de los algoritmos más sencillos y aplicados[29-31] a problemas de regresión y clasificación. Es un método de aprendizaje no supervisado y utiliza estimadores no paramétricos, es decir que son modelos no teóricos porque no buscan parámetros.

El algoritmo KNN predice patrones de comportamientos de datos nuevos (llamados conjunto de prueba) a partir de un conjunto de datos conocidos llamados conjunto de entrenamiento en la forma $(\vec{x} \mid y)$. Dado un valor de K, un conjunto de entrenamiento y un nuevo punto a predecir $\vec{x}_0$, el algoritmo primero identifica en el set de entrenamiento las K observaciones más cercanas a $\vec{x}_0$, este conjunto lo representaremos como A. Luego se estima el valor de la variable independiente en el punto $\vec{x}_0$ como el promedio de los valores de y de las observaciones en el conjunto A (ecuación (0.4)).

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x_i \in A} y_i \qquad (0.4)$$

En este método existen dos factores fundamentales que rigen su desempeño:

- El primero es la medida de similaridad utilizada para la selección de los K vecinos más cercanos. La más utilizada y prácticamente estándar es la distancia Euclidiana, pero también pueden ser usadas otras como la distancia Chebychev y la distancia Manhattan[32].
- El segundo parámetro que también afecta su desempeño es el valor de K. Mientras más bajo más flexible será el modelo y viceversa. Un modelo demasiado flexible (K) perderá información de las posibles tendencias, ya que el valor de y será promediado con muy pocos elementos (será muy local) teniendo un valor predictivo escaso. Un modelo demasiado rígido promediará los valores de y con demasiados elementos y por tanto se perderá información de las características locales. El valor de K debe ser entonces ajustado haciendo uso de un set de pruebas y observando el comportamiento del error vs K.

Árboles de regresión y Random Forest

Los árboles son una estructura de datos que han sido muy utilizados en la literatura. Estos tienen el atractivo de que sus modelos se asemejan mucho al proceso de toma de decisiones de un humano, y por tanto se encuentran entre los más fáciles de entender. Existen varios algoritmos que utilizan árboles para el análisis de datos, como ejemplos podemos citar los árboles de decisión (en sus variantes de árboles de regresión y árboles de clasificación) y los algoritmos genéticos.

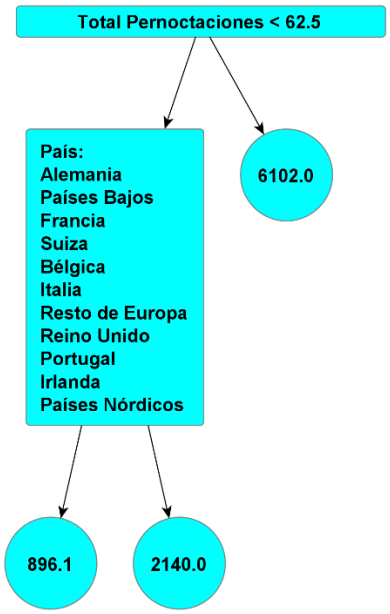

Figura 1: Ejemplo de un árbol de regresión de tres niveles obtenido de la data EGATUR. Primero se dividen los datos de acuerdo al número de pernoctaciones (izquierda los datos que cumplen con el criterio, derecha los que no), los datos que no cumplen con el criterio se les asigna el valor predicho 6102.0. Los datos que cumplen con el criterio de pernoctaciones son particionados de manera similar según si pertenecen o no a un grupo de países.

Los árboles de regresión, desarrollados en 1963 por Morgan y Sonquist[33], particionan un conjunto de datos en subconjuntos siguiendo una serie de reglas que actúan recursivamente en una data. Un árbol de ejemplo y muestra en la Figura 1. Cada regla de selección divide los datos en conjuntos cada vez más pequeños de manera que se vaya minimizando la suma de residuos cuadrados. Luego el valor que el árbol predice para cada observación en una región

determinada por estas reglas (nodos terminales) será el promedio de los valores de respuesta de las observaciones de entrenamiento que caen en esta región. Esta aproximación va a tener siempre una gran varianza, aunque por el otro lado, su gran sencillez y lo intuitivo de su interpretación lo vuelven una herramienta muy valiosa para identificar los factores más importantes que influyen en la variable respuesta.

Más recientemente, en la década de 1990 Kam desarrolló el algoritmo de Random Forest[34] (RF) muy relacionado con los árboles, el cual es una de las técnicas más utilizadas en la actualidad en Machine Learning por su gran flexibilidad y aplicabilidad a diferentes tipos de problemas[35-37].

Este algoritmo utiliza el hecho de que si se promedian n diferentes observaciones independientes de varianza σ, la varianza para este promedio es $\sigma/\sqrt{\mathrm{n}}$. De esta forma, la idea central de este algoritmo es que en lugar de crear un árbol para predecir la variable deseada, se escogen n subconjuntos del set de entrenamiento, con cada uno se crea un árbol diferente y la predicción total será el promedio de las predicciones de estos árboles.

Si en nuestro set de entrenamiento tenemos un predictor muy fuerte, en la mayoría de nuestros árboles la primera regla de división estará dada por el mismo. Esto trae como consecuencia que la mayoría de nuestros árboles estarán correlacionados, no siendo óptima la reducción de la varianza al promediar. Para evitar este problema, en la creación de cada uno de los árboles, no se utilizan todos los descriptores, sino un subconjunto de estos seleccionado aleatoriamente. De forma general, el número de descriptores a utilizar m es fijado como $m \approx \sqrt{d}$ siendo d el número total de predictores de la data.

Otro parámetro que afecta el comportamiento del modelo obtenido es el número de árboles a promediar. Si se utiliza un número demasiado alto, el modelo será muy flexible, ajustando perfectamente la data de entrenamiento perdiéndose así información de posibles tendencias (valor predictivo limitado), en caso de que el número sea demasiado bajo, el modelo será muy rígido, se tendrá información de las tendencias pero también una alta varianza.

Redes Neuronales y Aprendizaje Profundo

Las primeras investigaciones sobre las redes neuronales datan de 1958 cuando Frank Rosenblatt[38] creó la primera idea del perceptrón basado en el funcionamiento de una neurona biológica. Más tarde, en el año 1986 las primeras neuronas artificiales fueron modeladas por Rumelhart y McClellan[39].

Las redes neuronales procesan datos del pasado encontrando tendencias al futuro comportándose como un modelo muy simplificado de un cerebro biológico. Una red neuronal está compuesta de tres capas: la capa de entrada, la capa oculta y la capa de salida (ver Figura 2). La capa de entrada alimenta de datos a la capa oculta. La capa oculta está constituida por neuronas artificiales (funciones complejas que crean predictores nuevos) interconectadas que modifican los datos y los envían a la capa de salida. Por último la capa de salida, con las predicciones hechas por la capa oculta, produce el resultado final.

En la capa oculta cada neurona artificial imita a una neurona biológica. Cada neurona tiene de entrada un conjunto de predictores $\{x_i\}$, cada uno con un peso asociado, los que se emplean en una suma ponderada $sum = \sum_i w_i x_i$, al valor de esta suma se le aplica una función de paso $y = f(\mathrm{sum})$ y el valor obtenido de esta nos da la activación o salida de esta neurona. El procedimiento de entrenamiento consiste entonces en ajustar estos pesos utilizando los datos de entrenamiento de forma tal que para el conjunto de predictores de entrada, la función de respuesta dé lo más cercana a la real. Es decir, imitando el comportamiento de un cerebro biológico, la red neuronal aprende con ejemplos pasados y luego puede hacer predicciones.

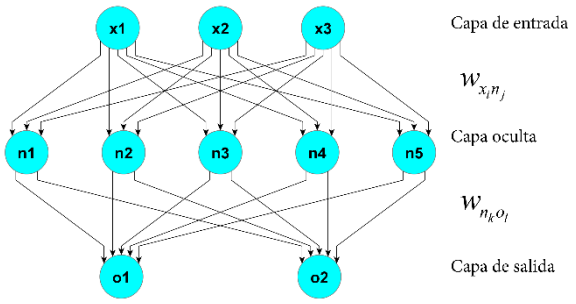

Figura 2: Esquema simplificado de una red neuronal, los datos de entrada.

El caso del llamado aprendizaje profundo o "Deep Learning" (DL) es un algoritmo basados en las ideas de las redes neuronales. A diferencia de una red neuronal, en el caso de DL, se utilizan varias capas ocultas. El proceso de obtención de un resultado entonces varía en que, una vez activadas las neuronas de la primera capa oculta, cada una envía sus valores a cada una de las neuronas de la segunda capa oculta, en la cual cada neurona pondera las señales utilizando otro set de pesos y que se manda a cada una de las neuronas de la siguiente capa oculta y así sucesivamente hasta que la última capa oculta envía los resultados a la capa de salida. Este procedimiento es más flexible que las redes neuronales de una sola capa oculta y ha sido utilizado con mucho éxito.

Máquinas de Vectores de Soporte (SVM)

Las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) son técnicas de aprendizaje

supervisado creadas en la década de 1990 por Vapnik y colaboradores[40] aunque la idea central del algoritmo se venía investigando desde antes[41]. Las SVM son otro de los algoritmos más aplicados en la actualidad para problemas de clasificación[42], regresión[43] y agrupamiento[44] pues son robustos en los casos en que hay una alta dimensionalidad.

La idea básica de esta técnica, cuando es utilizada para clasificación, se basa en el concepto del clasificador de máximo margen[45], el objetivo es buscar un el hiperplano desde el cual los puntos más cercanos al mismo tengan la máxima distancia del mismo posible. El algoritmo SVM basado en esta idea, optimiza el hiperplano, pero en lugar de buscar un hiperplano que separe las dos clases (el cual en muchos casos no existe), permite la presencia de puntos mal clasificados pero los penaliza en la optimización. También, para hacer el problema más separable, el algoritmo utiliza kernels, que son transformaciones matemáticas del espacio de variables, estas aumentan la dimensionalidad del espacio y hacen más separables las clases.

Cuando es utilizada para regresión, la técnica solo varía en el objetivo, que en lugar de buscar el hiperplano que mejor separe las clases, se busca el que más cerca tenga todas las observaciones[45]. En este caso lo que se va a penalizar es la distancia de cada observación del hiperplano, y los parámetros se ajustan disminuyendo esta.

Validación Cruzada

La validación cruzada (CV por sus siglas en inglés) es un método muy utilizado para estimar las capacidades predictivas de un algoritmo cuando se está aplicando a un problema. Se emplea con algoritmos de aprendizaje supervisado y no supervisado y estima las precisiones de los modelos que se vayan a utilizar con las bases de datos. Sus primeas aplicaciones se remontan a la década de 1960[46].

El método de validación cruzada de n iteraciones puede explicarse de la siguiente forma. Primero permutamos toda nuestra data. Posteriormente dividimos las observaciones en n conjuntos de aproximadamente igual tamaño. Luego, se selecciona uno de los grupos de datos como conjunto de prueba. Se entrena el algoritmo evaluado utilizando todos los datos que no pertenecen al conjunto seleccionado, y por último, se evalúan los parámetros predictivos (como suelen ser RMSE, R^2, MAE entre otros) de nuestro modelo entrenado utilizando el conjunto de prueba. Esto se repite para cada uno de los n grupos y posteriormente se busca la media aritmética o el promedio de esos valores de cada iteración en el conjunto de datos.

2. Materiales y Métodos

Los datos de los turistas fueron obtenidos de la encuesta de gasto turístico (EGATUR) del instituto nacional de estadística de España[26] en el periodo de tiempo comprendido entre octubre de 2015 y agosto de 2017. Estos fueron analizados empleando los algoritmos LR, KNN, RF, DL y SVM. Todos los cálculos fueron ejecutados utilizando el paquete de software estadístico R[47]). Como manejador de los algoritmos se utilizó la librería MLR[48]. Los métodos LR, KNN, RF, DL y SVM se utilizaron como están implementados en los paquetes stats[49], kknn[50], randomForest[51], h2o.deeplearning y libsvm[52] respectivamente del repositorio de R.

Para el análisis de cada algoritmo se procedió en tres fases. Primero se analiza la dependencia de la exactitud de las predicciones y los hiperparámetros fundamentales para cada algoritmo. Una vez seleccionado los valores óptimos de los mismos, se estiman el RMSE, MAE y R^2 para las predicciones utilizando validación cruzada de 10 iteraciones.

Por último se analiza el comportamiento de los residuos para observar la presencia de tendencias del modelo a desviarse de los datos experimentales. Para esto se utiliza el método del set de validación donde se divide la data de forma aleatoria en dos subconjuntos de observaciones, entrenamiento y prueba a una razón del número de datos de 0.75/0.25 respectivamente. Utilizando los datos de entrenamiento se entrena el modelo y se mide su capacidad predictiva con el set de prueba. En este caso se predijo el gasto total para cada una de las observaciones del set de prueba y de entrenamiento. Esto nos permite obtener los residuos del modelo para ambos conjuntos de datos.

3. Resultados y discusión

Preparación de la data

Primeramente se procedió a la preparación de la data a utilizar. La encuesta EGATUR para el periodo comprendido entre octubre de 2015 a agosto de 2017 tiene 173145 observaciones. En cada uno de los casos se obtienen 49 descriptores los cuales son: fecha, identificador del cuestionario, pernoctaciones etapa 1, … , pernoctaciones etapa 18, número total de pernoctaciones, comunidad autónoma etapa 1, … , comunidad autónoma etapa 18, encuesta, TEN, vía de salida del país, país de origen del turista, comunidad autónoma donde estuvo la mayor parte de su estancia, el número total de pernoctaciones, el tipo de alojamiento, el motivo del viaje, paquete turístico y el gasto total, la documentación más detallada del significado y características de cada descriptor puede obtenerse en la documentación[53].

El descriptor de fecha fue dividido en dos descriptores, mes y año. En este trabajo solo se tienen datos de

2 años aproximadamente, pocos años para que una tendencia al crecimiento o decrecimiento debido al año sea extrapolable analizando la data como una serie temporal. Debido a esto, el descriptor de "año" fue descartado.

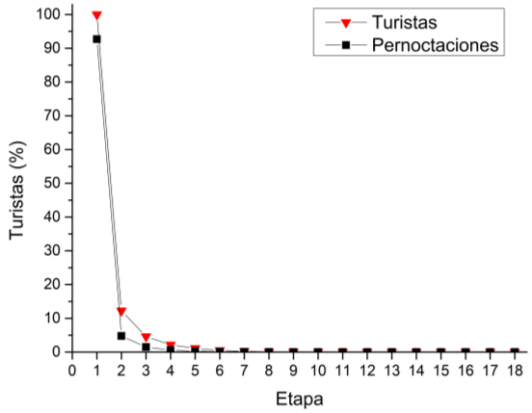

Figura 3: Porcentaje del total de turistas con cada número de etapas en su viaje (triángulo rojo) y porcentaje del número total de pernoctaciones que se hace en cada etapa (cuadrado negro).

En el caso de los descriptores "identificador del cuestionario" y "procedencia de la encuesta", los mismos no afectan el comportamiento de los turistas y por tanto su gasto, ya que son creados solo para seguir la pista de los datos y su procedencia cuando se confecciona la base de datos, por lo tanto fueron desechados.

Otro descriptor que fue descartado es el TEN, este separa al turista no residente (no en tránsito) del turista residente en tránsito. En este caso el 94.8% de las observaciones se trataban de turistas no residentes (no en tránsito), que representa prácticamente la totalidad, siendo este un descriptor que diferencia solo a una pequeña minoría de los datos.

La Figura 3 muestra para cada etapa, qué por ciento del total de turistas estuvo y que por ciento del total de pernoctaciones en todas las observaciones se efectuaron en cada etapa. Es posible observar que de los turistas, solo una pequeña minoría, menor del 13 por ciento participa en más de una etapa en su viaje turístico, solo un 12.26% tiene un viaje de al menos dos etapas, y para más etapas el número es aún menor. El comportamiento de las pernoctaciones en cada etapa es similar, el 92.67% de ellas se realiza en la primera etapa. Debido a esto, es de esperar que solo sea necesario tener en cuenta la primera etapa para el número de pernoctaciones y la comunidad autónoma del viaje. Otra aproximación más sutil es tomar los datos para cada turista del número total de pernoctaciones de la comunidad autónoma donde más pernoctaciones realizó, que son otros descriptores dados por la encuesta, de esta forma se eliminan los descriptores que desglosan las pernoctaciones y comunidades autónomas por etapas.

De esta forma la data utilizada en este trabajo queda con solo 9 descriptores: "mes"; "vía de salida"; "país de origen", "comunidad autónoma"; "pernoctaciones totales"; "alojamiento"; "motivo"; "paquete turístico" y gasto total.

Regresión lineal

Este método es muy sencillo y no presenta hiperparámetros, entonces se procedió a estimar el desempeño del mismo para predecir el gasto total de un turista de acuerdo a esta data. En este caso se utilizó una estrategia de muestreo de validación cruzada de 10 iteraciones, de la cual se extrajeron varias medidas de la varianza del método en el set de prueba y el de entrenamiento. Esto nos permite estimar cuánta será la varianza del método en sus predicciones. Los resultados promedios para el set de prueba fueron 268.0 para el error absoluto medio (MAE), 579.9 para el error cuadrático medio (RMSE) y en el caso del R^2 se obtiene un promedio de 0.698. Para el set de entrenamiento, los resultados fueron similares ($MAE = 268.0$, $RMSE = 579.2$, $R^2 = 0.70$), indicando ausencia de overfitting. Esto es esperado, ya que este es uno de los modelos más simples, que captura el carácter de las relaciones entre descriptores de forma sencilla y confiable pero presenta una alta varianza. De esta forma se observa un relativamente bajo valor del R^2 debido a la incapacidad del método de reproducir dependencias no lineales.

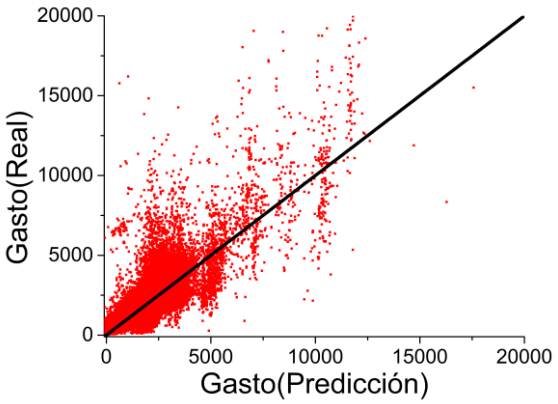

Figura 4: Dispersión de los valores del gasto total de los turistas (puntos rojos) respecto al gasto predicho por la regresión lineal, línea negra.

Al realizar un ajuste lineal, esta vez utilizando toda la data se obtuvieron los parámetros mostrados en la Tabla 1 del Apéndice A. Todas las variables son estadísticamente significativas. Puede observarse que en el caso de la variable de "Comunidad autónoma", algunas comunidades tienen un valor de $p > 0.05$, esto no es indicativo de que la variable de "Comunidad autónoma" no es significativa, dado que la variable es de niveles, esto solo indica que este nivel (es

decir esta comunidad autónoma) no es estadísticamente diferente del nivel de referencia (con coeficiente 0), que en este caso es la comunidad autónoma de Madrid. Más detalles acerca del funcionamiento de la regresión lineal para variables de niveles pueden obtenerse en la bibliografía[45].

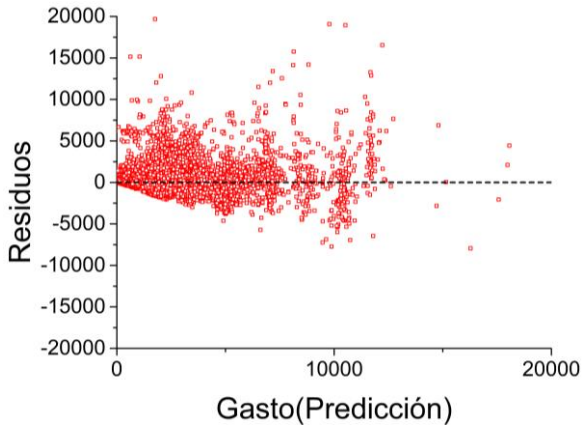

Figura 5: Residuos del ajuste lineal del gasto total.

La Figura 5 muestra los residuos obtenidos de este ajuste, no obstante por su alta dispersión, estos parecen estar distribuidos aleatoriamente centrados en el 0, indicando ausencia de posibles tendencias que nuestro ajuste no haya tenido en cuenta. Al realizar pruebas utilizando términos de 2 o de interacción se produce un incremento en el R^2 y disminución del MAE y RMSE pero con una afectación marginal.

K vecinos más cercanos

En el caso del algoritmo KNN se obtuvo el valor de K óptimo para hacer predicciones. Para esto se realizaron ajustes utilizando un muestreo de validación cruzada de 5 iteraciones para varios valores de K en el rango de 2 a 60 buscando el valor de K, que minimice para el set de prueba los valores de MAE y RMSE promedios y que maximice el R^2.

La Figura 6 muestra los resultados del MAE promedio vs el valor de K utilizado. La Figura 20 del Apéndice B muestra el conjunto de las otras medidas de calidad del ajuste (RMSE y R^2) para cada K.

Puede observarse que la zona donde se localiza la K óptima se encuentra en el rango de valores entre 8 y 10 de acuerdo al MAE. Para el caso de $K = 6$ se observa un mínimo pronunciado. Este, aun cuando presenta un menor MAE que todos los demás valores de K, y se observa el gráfico correspondiente al RMSE este valor no cae cerca del mínimo de la curva. Similar comportamiento se observa para el R^2, donde para $K = 6$ no se está cerca del máximo. Si observamos los gráficos correspondientes a RMSE y R^2 podemos observar que según estos criterios (el RMSE mínimo y el R^2 máximo), la zona de $8 \le K \le 10$ está dentro de la zona óptima para estas otras medidas de bondad de ajuste. Debido a estas razones, se tomó 9 como valor de K óptimo, pues es un buen compromiso para todas las medidas del ajuste ($MAE = 251.3$, $RMSE = 570.8$ y R^2=0.706).

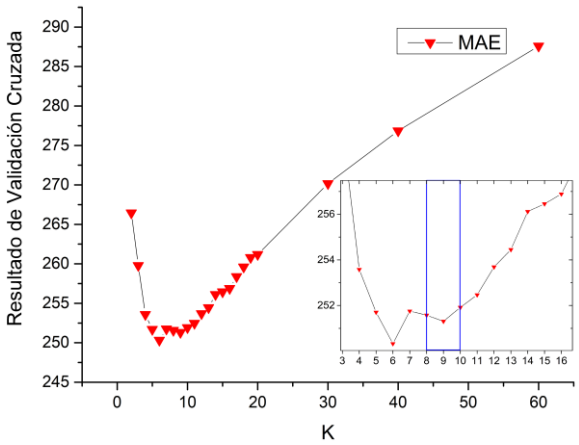

Figura 6: Dependencia del error absoluto promedio (MAE) vs k utilizado en KNN para el set de prueba. Para cada k se reporta un promedio del MAE de 10 iteraciones de validación cruzada. Se muestra ampliada la zona cercana al mínimo de la curva.

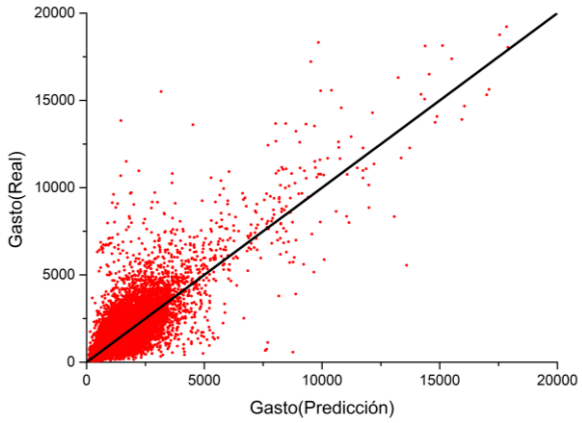

Figura 7: Gasto total real de los turistas vs Gasto predicho para KNN con $K = 9$. La línea negra muestra las predicciones del modelo, los puntos rojos los valores reales.

La metodología del set de validación, empleando $K = 9$ reporta para el set de prueba resultados muy similares a la CV. El MAE fue 254.2, el RMSE 567.5 mientras que para el R^2 su valor fue de 0.70. La Figura 7 da una medida del comportamiento en el set de prueba de lo obtenido. Puede apreciarse que los valores reales, representados por los puntos rojos están a ambos lados del modelo (línea negra) y centrados en el mismo. Esto es indicativo de que el modelo captura la tendencia de los datos. Esto se

corrobora en la Figura 8 que muestra los residuos del ajuste, tanto para el set de prueba como el de entrenamiento. Aquí puede observarse que para ambos conjuntos los residuos están uniformemente distribuidos de forma aleatoria centrados en el 0.

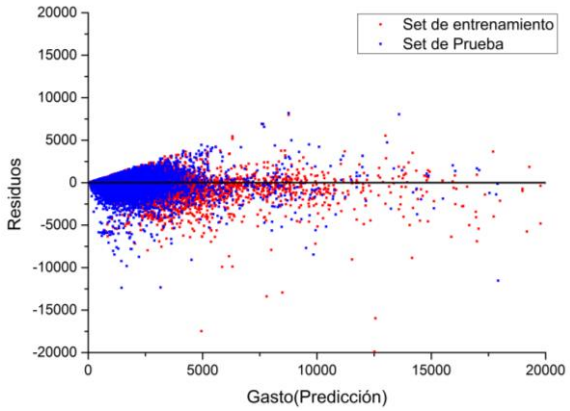

Figura 8: Residuos de los ajustes empleando el método KNN($K = 9$).

Random Forest

En el caso de este algoritmo, para su ajuste se pueden cambiar dos hiperparámetros; el número de descriptores a usar en cada árbol y el número de árboles que se utilizan en el conjunto. En el caso del número de descriptores a usarse en cada árbol, se dejó como la raíz cuadrada del número total de descriptores pues es una práctica muy utilizada en la literatura con buenos resultados[45].

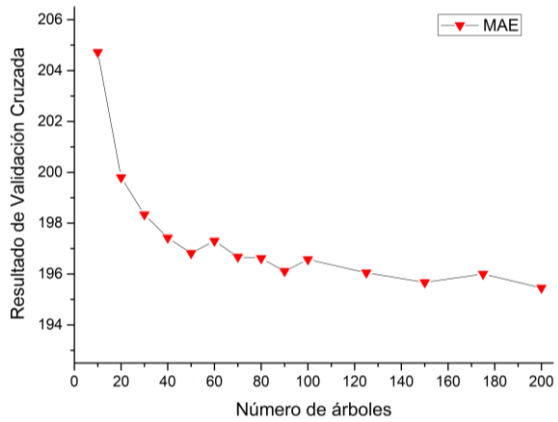

Figura 9: Dependencia del error absoluto promedio (MAE) vs el número de árboles usado en RF, para el set de prueba. Para valor del número de árboles se reporta un promedio del MAE de 10 iteraciones de validación cruzada.

Para encontrar el número de árboles óptimos a utilizar se estimaron el MAE, RMSE y R^2 para varios valores del número de árboles en el rango de 10 a 200. Los resultados para cada valor del número de árboles se estimaron utilizando validación cruzada de 10 iteraciones.

La Figura 9 muestra la dependencia del MAE con el número de árboles empleados. La Figura 21 del Apéndice B muestra gráficos similares para el RMSE y R^2. Puede observarse un rápido declive del MAE con el número de árboles y posteriormente un comportamiento casi asintótico. Teniendo en cuenta que el incremento del número de árboles conlleva a un incremento del costo computacional, se escogió un valor del número de árboles en la zona donde la curva deja de disminuir rápidamente, en este gráfico esta zona puede observarse para un número de árboles entre 50 y 100. En el caso de los gráficos de RMSE y R^2 (Figura 21 del Apéndice B) la zona de cambio de pendiente correspondiente se encuentra sobre el rango de 80-100, teniendo en cuenta esto, se escogió 80 como el valor óptimo del número de árboles. Este se corresponde con unos valores de $MAE = 196.6$, $RMSE = 474.9$ y $R^2 = 0.80$, ligeramente superiores a los resultados obtenidos para KNN y LR.

Al utilizar el método del set de validación utilizando 80 árboles los resultados obtenidos para el set de prueba fueron muy cercanos al resultado obtenido por la validación cruzada ($MAE = 198.5$, $RMSE = 470.9$, $R^2 = 0.80$) y ligeramente superiores al obtenido por KNN. También, los valores reales en el set de prueba del gasto total están distribuidos uniformemente por encima y debajo de las predicciones de RF (Figura 10)

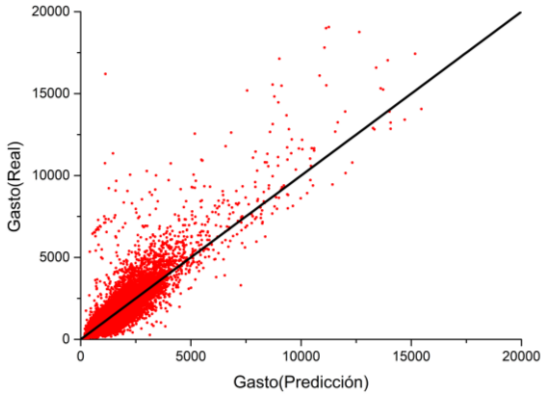

Figura 10: Gasto total real de los turistas vs Gasto predicho. La línea negra muestra las predicciones del modelo RF con 80 árboles, los puntos rojos los valores reales.

Esto indica que el modelo obtenido no tiende a sobreestimar o subestimar los valores reales, capturando de forma correcta las posibles tendencias en los datos. El análisis de los residuos muestra resultados similares, Figura 11, donde se puede observar tanto para el set de entrenamiento como el de prueba que los residuos están uniformemente distribuidos por encima y por debajo del 0 corroborando la ausencia de tendencias

de las predicciones del modelo a alejarse de la data. También se evidencia que las amplitudes de las distribuciones de los residuos son similares para el set de entrenamiento y el de prueba, indicando la ausencia de sobreajuste.

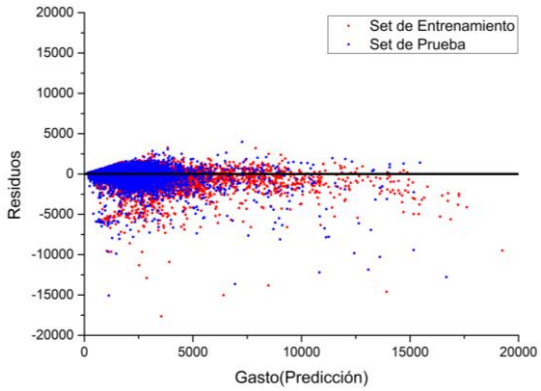

Figura 11: Residuos de los ajustes empleando el método RF (80 árboles).

Máquinas de Vectores de Soporte (SVM)

En el caso del algoritmo SVM se utilizó la variante de regresión $\varepsilon - SVM$, aquí el valor de ε define la distancia que los puntos pueden alejarse del hiperplano sin que se penalice la función a minimizar. En los casos en que se alejen más allá del margen definido por ε, estos se penalizan de forma lineal con una pendiente, que es un parámetro a definir llamado costo. En este caso el kernel utilizado fue el radial.

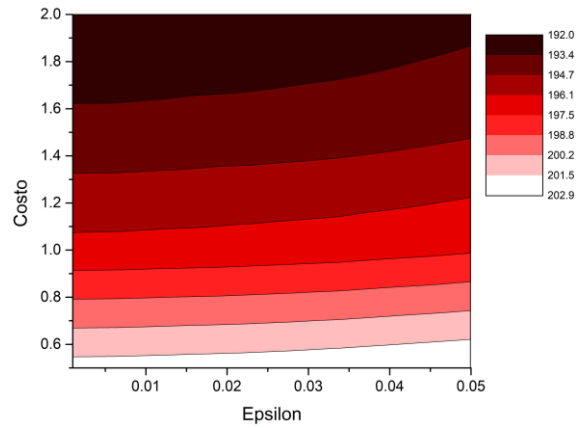

Figura 12: Dependencia del MAE del SVM vs los parámetros costo y épsilon.

Para encontrar valores adecuados para el costo y ε, se utilizó el método de set de validación dividiendo la data en entrenamiento y prueba (en relación de observaciones 0.75/0.25 respectivamente). Se realizó una malla de posibles valores del costo y ε. Para el costo se utilizaron los valores en el rango 0.5-2.0 con una diferencia entre dos valores consecutivos de 0.5 (4 valores posibles). En el caso de ε se utilizó el rango de 0.001 a 0.05, donde la diferencia entre dos puntos consecutivos fue $5.44 \cdot 10^{-3}$ (10 valores posibles). Para cada combinación posible de costo y ε se entrenó el SVM usando el set de entrenamiento y con este, se predijeron los valores para el set de prueba, obteniéndose el MAE, RMSE y R^2. El número total de combinaciones de costo y ε a calcular fue 40 y dado el costo computacional de cada punto, se hace imprácticco usar una estrategia de validación cruzada, por lo que solo se utilizaron los estimados de este par entrenamiento/prueba para cada caso.

La Figura 12 muestra la dependencia del MAE con el costo y ε. Puede observarse la tendencia a disminuir el MAE a mayores valores de costo y menores valores de ε. El conjunto óptimo de parámetros es $\varepsilon = 0.001$ y costo $= 2$, con valores de $MAE = 192.0$, $RMSE = 516.5$ y $R^2 = 0.77$ para el set de prueba. El hecho de que los parámetros óptimos encontrados se encuentren en la frontera del rango escaneado, parece indicar que los valores óptimos de costo y ε se encuentran fuera del mismo. No obstante los resultados de MAE, RMSD y R^2 obtenidos para $\varepsilon = 0.001$ y costo $= 2$ son muy similares a los obtenidos por KNN y RF por lo que son adecuados para este trabajo. La disminución esperada con una búsqueda más exhaustiva difícilmente justificaría el costo computacional necesario.

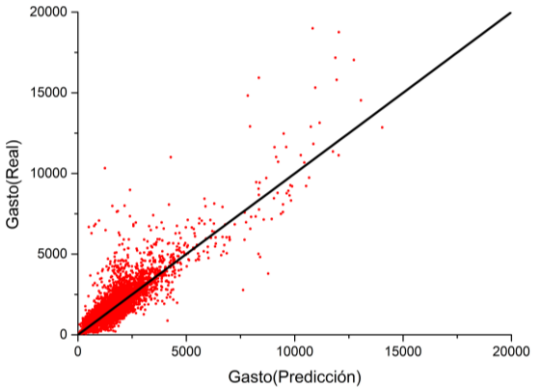

Figura 13: Gasto total real de los turistas vs Gasto predicho. La línea negra muestra las predicciones del modelo SVM ($\varepsilon = 0.001$ **y** costo $= 2$), los puntos rojos los valores reales.

Para una mejor estimación de las capacidades predictivas del método con el ε y costo seleccionados, se realizó una validación cruzada de 10 iteraciones en este punto. Los resultados promedios obtenidos fueron $MAE = 188.9$, $RMSE = 497.3$ y $R^2 = 0.78$ para los sets de prueba, sin diferencias significativas respecto a los resultados preliminares con un solo par entrenamiento/prueba.

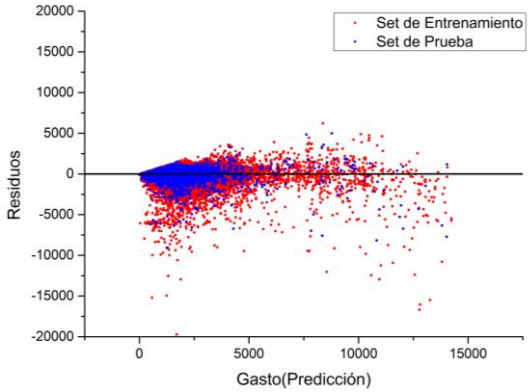

Figura 14: Residuos de los ajustes empleando el método SVM ($\varepsilon = 0.001$ y costo $= 2$).

Para la metodología del set de validación los resultados fueron ($MAE = 190.5$, $RMSE = 572.8$ y $R^2 = 0.72$). La Figura 13 muestra la distribución del gasto real respecto a las predicciones del modelo obtenido. Es posible observar que el modelo captura las tendencias en los datos de los turistas. De igual forma la Figura 14 muestra los residuos tanto para el set de prueba como para el de entrenamiento. Ambos gráficos muestran que el método SVM con los parámetros seleccionados no presenta ninguna tendencia en los residuos, ya que se encuentran distribuidos aleatoriamente por encima y debajo del 0. De igual manera la similitud en las distribuciones de los residuos para el set de entrenamiento y el de prueba evidencia la ausencia de sobreajuste.

Aprendizaje profundo

El algoritmo de aprendizaje profundo, en su implementación del paquete h2o.deeplearning, depende de varios hiperparámetros. Los dos hiperparámetros más influyentes son el número de capas y el número de neuronas en cada capa. La Figura 15 y la Figura 16 muestran la dependencia del MAE y RMSE respectivamente con estos parámetros.

Tanto las curvas de nivel del MAE como del RMSE muestran varios mínimos. El objetivo es encontrar una combinación de valores de hiperparámetros que minimice tanto el MAE como el RMSE. Se seleccionó la combinación de 170 capas de neuronas con 230 neuronas cada una pues en ambos gráficos corresponde a un mínimo.

Al realizar una validación cruzada de 10 iteraciones, los resultados promedios para el set de prueba fueron: $MAE = 191.1$, $RMSE = 454.2$ y $R^2 = 0.81$. Con un valor ligeramente superior del MAE respecto al SVM, no siendo así con el RMSE, donde el desempeño de DL es superior. Estos resultados indican que la exactitud de los dos métodos son comparables para esta data.

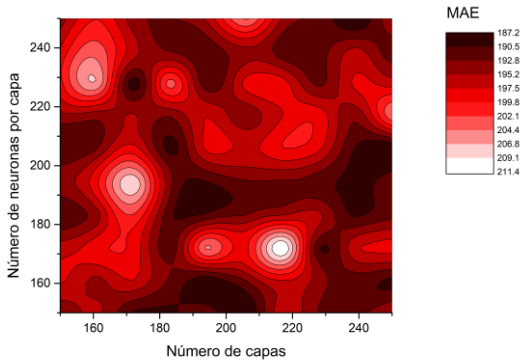

Figura 15: Dependencia del MAE en DL con el número de capas y el número de neuronas en cada una.

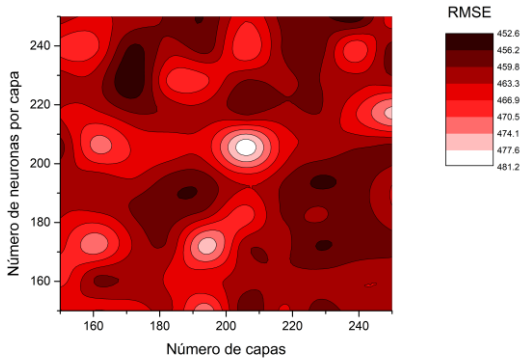

Figura 16: Dependencia del RMSE en DL con el número de capas y el número de neuronas en cada una.

Al utilizar la metodología del set de validación, los resultados fueron: $MAE = 189.9$, $RMSE = 449.2$ y $R^2 = 0.81$. El comportamiento de los valores reales del gasto en comparación con las predicciones del modelo puede observarse en la Figura 17. Es posible observar que los valores de gasto real de los turistas están distribuidos en una franja aproximadamente por el centro de la cual pasa nuestro modelo. No se observa evidencia de que el modelo sobrestime o subestime los datos experimentales así como no se observa la presencia de alguna tendencia del modelo a desviarse de los datos. Esto también puede verificarse observando el comportamiento de los residuos Figura 18. Aquí se observa que tanto los residuos de los datos usados para el entrenamiento como los datos de prueba están distribuidos de forma aleatoria centrados en el 0. También puede observarse que ambas distribuciones se superponen dentro de los mismos márgenes lo cual indica que en este caso no hay un alto grado de "overfitting", problema muy común en los modelos que implican redes neuronales.

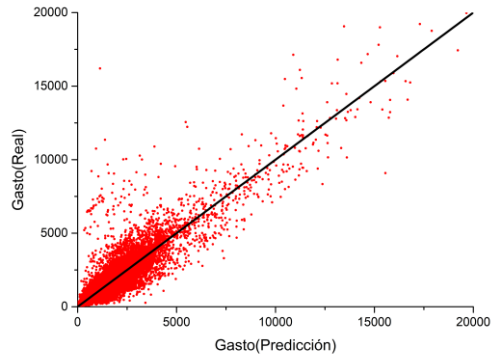

Figura 17: Gasto total real de los turistas vs Gasto predicho. La línea negra muestra las predicciones del modelo DL, los puntos rojos los valores reales.

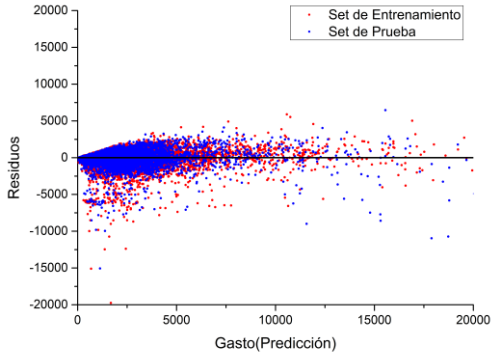

Figura 18: Residuos de los ajustes empleando el método DL.

Análisis de correlación entre los modelos

En modelos no correlacionados, los residuos de los mismos serán aleatorios. Debido a esto al promediarlos se produce una cancelación de errores y una disminución de la varianza.

Para analizar la factibilidad de esta estrategia se dividió la data total en 6 subconjuntos de aproximadamente igual tamaño de observaciones escogidas aleatoriamente. Se procedió a entrenar cada uno de los 5 métodos usados cada uno con un subconjunto diferente. El grupo restante de datos se utilizó como set de prueba y con cada método se realizaron predicciones y se encontraron los residuos para estas predicciones con cada método. Se buscó entonces el valor de la correlación de los residuos para cada par de métodos.

Se encontró que todos los modelos utilizados tienen una alta correlación en sus errores, en el rango de 0.68-0.89, ver Tabla 2. El alto valor de correlación de los residuos implica que promediar los modelos no traería ningún beneficio, pues la cancelación de errores depende de la distribución aleatoria de los mismos en cada método, por lo que no sería efectiva. Debido a esto, para modelar el gasto de los turistas, es recomendable utilizar en este caso el método que mejor relación precisión/costo nos ofrece, que en este caso es el DL.

Análisis de la importancia de variables

Con el objetivo de encontrar los factores que más influyen en el gasto de un turista se procedió a analizar la importancia relativa de cada variable. Para esto se utilizó la metodología del set de validación entrenando los cinco métodos estudiados utilizando sus hiperparámetos óptimos encontrados previamente.

Una vez obtenidos los modelos se procedió a cuantificar la importancia de cada variable para cada modelo. Para esto, se permutan aleatoriamente los valores de la columna correspondiente a la variable que queremos ver su impacto. Se procede entonces con cada método a predecir el gasto total utilizando la data con esta columna modificada y se halla el RMSE. El aumento del RMSE producto de estas permutaciones nos da una medida de la importancia relativa de la variable para cada método. Mientras más importante es una variable para un modelo, mayor dependerán los resultados del mismo del valor de esta y por tanto mayor será el aumento del RMSE del mismo si se permutan los valores de esta variable. Para medir este aumento, puede usarse como referencia el promedio de los valores de RMSE obtenidos para cada método utilizando validación cruzada con 10 ($\overline{RMSE} = 514.8$).

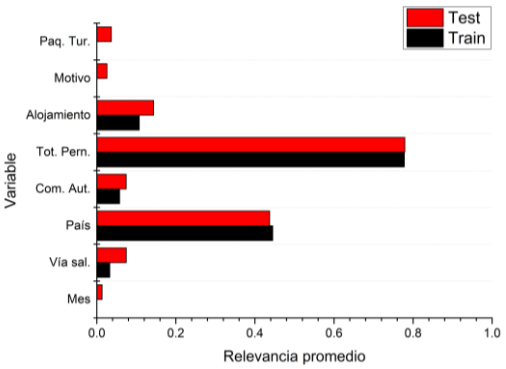

Figura 19: Relevancia de cada variable obtenida como el promedio para los métodos utilizados.

La Figura 19 muestra las relevancias relativas de cada variable como promedio de los diferentes métodos. Para hallarlas, los aumentos de RMSE obtenidos para cada método fueron promediados y los valores obtenidos escalados de 0 a 1. La variable más importante fue el número total de pernoctaciones, con un índice de 0.78 para el set de prueba, le sigue el país de origen (0.43) y con menos relevancia continúan el tipo de alojamiento (0.14), la comunidad autónoma (0.07) y la vía de salida (0.07). Los valores de la

importancia relativa de cada variable para cada método se muestran en la Tabla 3 del Apéndice A.
Si además se observan los valores de las pendientes, obtenidos previamente en la regresión lineal (Tabla 1, Apéndice A), es posible observar que además de ser la variable más importante, el número de pernoctaciones tiene un efecto positivo sobre el gasto.
En el caso de la segunda variable más importante, esta resultó ser el país de procedencia. El análisis de las pendientes para los diferentes niveles de esta variable nos muestra que el gasto de los turistas respecto al país de procedencia se ordena de la siguiente forma: Resto del Mundo > Resto de América > EEUU > Rusia > Países Nórdicos > Países Bajos > Resto de Europa > Bélgica > Irlanda > Alemania > Suiza > Francia > Reino Unido > Portugal > Italia.
De estos resultados podemos concluir que para maximizar el gasto de los turistas y por tanto aumentar las ganancias de la industria turística, se deben aumentar las ofertas que incrementen el número de pernoctaciones de los turistas, especialmente las dedicadas a los turistas provenientes de países que se incluyen en el grupo "Resto del Mundo".

4. Conclusiones

Se entrenaron los modelos LR, SVM, DL, RF y KNN utilizando la data de la encuesta del gasto turístico en España (EGATUR) entre octubre de 2015 y agosto de 2017 para predecir el gasto turístico. Todos los modelos tuvieron desempeños relativamente similares con MAE, RMSE y R^2 estimados en los rangos 188.9(SVM)-268.0(LR), 454.2(DL)-579.9(LR) y 0.698(LR)-0.810(DL) respectivamente. Los modelos se ordenan según el RMSE de sus predicciones de la siguiente manera DL<RF<SVM<KNN<LR.
Los métodos tienen una alta correlación en sus residuos, por lo que no es posible mejorar la capacidad predictiva de los mismos promediándolos. El mejor modelo predictivo en relación calidad/costo fue DL para el estudio realizado. Este presenta unos valores estimados por validación cruzada (10 iteraciones) de 191.1, 454.2 y 0.810 para sus MAE, RMSE y R^2 respectivamente. El SVM tiene un desempeño similar en cuanto a sus capacidades predictivas pero su costo computacional es varias veces mayor.
El análisis de la importancia de los diferentes descriptores arrojó que el principal factor en el gasto del turista es el número de pernoctaciones del mismo. Le sigue el país de procedencia y con menor importancia el alojamiento y la comunidad autónoma. Los otros descriptores presentan una relevancia marginal para predecir el gasto.

Referencias

(1) Kalinin, S. V.; Sumpter, B. G.; Archibald, R. K. *Nat. Mater.* **2015**, *14*,973–980.
(2) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. *Phys. Rev. Lett.* **2015**, *114*,105503
(3) Arsenault, L.-F.; Lopez-Bezanilla, A.; Lilienfeld, O. A. v.; Millis, A. J. *Phys. Rev. B* **2014**, *90*,155136
(4) Rubin, D. B. *Ann Intern Med.* **1997**, *127*, (8_Part_2),757-763.
(5) Kononenko, I. *Artificial Intelligence in Medicine* **2001**, *23* 89-109.
(6) Erevelles, S.; Fukawa, N.; Swayne, L. *Journal of Business Research* **2016**, *69*, (2),897-904.
(7) D. J. MacInnis, D. J. *The Journal of Marketing* **2011**, *75*, (4),136-154.
(8) Rabl, T.; Gómez-Villamor, S.; Sadoghi, M.; Muntés-Mulero, V.; Jacobsen, H.-A.; Mankovskii, S. *Proceedings of the VLDB Endowment* **2012**, *5*, (12),1724-1735.
(9) Akerkar, R. *International Journal of Computer Science and Applications* **2014**, *11*, (3),116-127.
(10) Lee, J.; Lapira, E.; Bagheri, B.; Kao, H.-a. *Manufacturing Letters* **2013**, *1*,38-41.
(11) Rodríguez-Mazahua, L.; Rodríguez-Enríquez, C.-A.; Sánchez-Cervantes, J. L.; Cervantes, J.; García-Alcaraz, J. L.; Alor-Hernández, G. *The Journal of Supercomputing* **2016**, *72*, (8),3073–3113.
(12) Schmidt, M.; Lipson, H. *Science* **2009**, *324*, (5923),43-44.
(13) Melnikov, A. A.; Nautrup, H. P.; Krenn, M.; Dunjko, V.; Tiersch, M.; Zeilinger, A.; Briegel, H. J. *Proceedings of the National Academy of Sciences* **2018**,DOI:10.1073/pnas.1714936115
(14) H, C.; RHL, C.; VC, S. *Manag Inf Syst Q (MIS) Q* **2012**, *36*, (4),1165-1188.
(15) Columbus, L. In *Forbes*; "Data Scientist Is the Best Job In America According Glassdoor's 2018 Rankings": https://www.forbes.com/sites/louiscolumbus/2018/01/29/data-scientist-is-the-best-job-in-america-according-glassdoors-2018-rankings/#45dc81435535, 2018.
(16) Tugcu, C. T. *Tourism Management* **2014**, *42*,207-212.
(17) Shakouri, B.; Yazdi, S.; Nategian, N.; Shikhrezaei, N. *J Tourism Hospit* **2017**, *6*, (4).
(18) Chen, C.-F.; Chiou-Wei, S. Z. *Tourism Management* **2009**, *30* 812–818.
(19) Eeckels, B.; Filis, G.; Leon, C. *Tourism Economics* **2012**, *18* (4),817–834.
(20) Massidda, C.; Mattana, P. *Journal of Travel Research* **2013**, *52*, (1),93–105.

(21) Durbarry, R. *Annals of Tourism Research* **2002**, *29*, (3),862–865.
(22) Khan, H.; C Seng; Cheong, W. *Annals of Tourism Research* **1990**, *17*,408–418.
(23) Marín, C. E.; Pérez, A. M. *Cuadernos de Turismo* **1998**, *2*,41-54.
(24) MINCOTUR España logra un récord de llegadas en 2017 con 82 millones de turistas internacionales; http://www.mincotur.gob.es/es-es/gabineteprensa/notasprensa/2017/documents/180110%20np%20balance%20turismo%202017.pdf; Last Update: 10/01/2018 Last Acces: 2/09/2018
(25) Tourism accounts for 16% of Spain's Gross Domestic Product, according to CaixaBank Research; https://www.caixabank.com/comunicacion/noticia/caixabank-research-study_en.html?id=40215; Madrid; Last Update: 07/06/2017 Last Acces: 10/09/2018
(26) *EGATUR Encuesta de gasto turístico*, Ministerio de Industria, Energía y Turismo, **2017**, Madrid, http://www.ine.es/daco/daco42/egatur/egatur1217.pdf .
(27) Aldrich, J. *Statistical Science* **2005**, *20*, (4),401–417.
(28) Fix, E.; Hodges, J. L. "Discriminatory analysis, nonparametric discrimination: Consistency properties.," USAF School of Aviation Medicine, 1951.
(29) Dhanabal, S.; Chandramathi, S. *International Journal of Computer Applications* **2011**, *31*, (7),14-22.
(30) Manikandan, R.; Sivakumar, R. *International Journal of Academic Research and Development* **2018**, *3*, (2),384-389.
(31) Victor, A.; Ghalib, M. R. *Research J. Pharm. and Tech* **2017**, *10*, (11),4093-4098.
(32) Mulak, P.; Talhar, N. *International Journal of Science and Research* **2015**, *4*, (7),2101-2104.
(33) Morgan, J. N.; Sonquist, J. A. *Journal of the American Statistical Association* **1963**, *58*,415-434.
(34) Ho, T. K. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal* **1995**,278–282.
(35) Belgiua, M.; Drăguţb, L. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *114*,24-31.
(36) Gao, D.; Zhang-H., Y. *Research in Astronomy and Astrophysics* **2009**, *9*, (2),14-39.
(37) Flaxman, A. D.; Vahdatpour, A.; Green, S.; James, S. L.; Murray, C. L. J. *Population Health Metrics* **2011**, *9*, (29),1-11.
(38) Rosenblatt, F. *Psychological Review* **1958**, *65*,386-408.
(39) Rumelhart, D.; McClelland, J. *Parallel Distributed Processing* **1986**, *1*,318-362.
(40) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*; ACM Press: New York, NY, USA, 1992, p 144–152.
(41) Vapnik, V.; Lerner, A. *Automation and Remote Control* **1963**, *24*,774–780.
(42) Bazi, Y.; Melgani, F. *IEEE Transactions on Geoscience and Remote Sensing* **2006**, *44*, (11),3374-3385.
(43) Flake, G. W.; Lawrence, S. *Machine Learning* **2002**, *46*, (1-3),271–290.
(44) Mozafari, A. S.; Jamzad, M. *Computer Vision and Image Understanding* **2017**, *162*,116-134.
(45) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, Heidelberg, Dordrecht, London, 2013.
(46) Mosteller, F.; Tukey, J. W. In *Handbook of Social Psychology*; Addison-Wesley: Reading, MA, 1968.
(47) R Development Core Team; R Foundation for Statistical Computing,: Vienna, Austria, 2008.
(48) Bischl, B.; Lang, M.; Kotthoff, L.; Schiffner, J.; Richter, J.; Studerus, E.; Casalicchio, G.; Jones, Z. M. *Journal of Machine Learning Research* **2016**, *17*, (170),1-5.
(49) Wilkinson, G. N.; Rogers, C. E. *Applied Statistics* **1973**, *22*,392–399.
(50) Samworth, R. J. *Annals of Statistics* **2012**, *40*,2733-2763.
(51) Breiman, L. *Machine Learning* **2001**, *45*, (1),5-32.
(52) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. ; http://www.csie.ntu.edu.tw/~cjlin/libsvm; Last Update; Last Acces:
(53) MINCOTUR DISEÑO DE REGISTRO FICHEROS DE DATOS ESTANDAR – EGATUR EXTENDIDO; http://estadisticas.tourspain.es/es-ES/estadisticas/egatur/microdatos/Documents/Dise%C3%B1o_Registro_Egatur.pdf; Madrid; Last Update: -; Last Acces: 3/05/2018

Apéndice A. Tablas

Tabla 1: Parámetros del ajuste para la regresión lineal.

Coeficiente	Valor	Error Estándar	Valor t	Valor p
Intercepto	149.4701692	8.2873	18.036	<2e-16
Tot. Pern.	54.09315774	0.1126	480.479	<2e-16
País:Resto del Mundo	1393.359582	7.1134	195.877	<2e-16
País:Resto de América	1256.456549	8.0281	156.508	<2e-16
País:EEUU	1085.33954	9.0818	119.507	<2e-16
País:Rusia	586.0461823	12.8814	45.495	<2e-16
País:Países Nórdicos	220.7127618	6.7135	32.876	<2e-16
País:Países Bajos	121.3819078	7.0607	17.191	<2e-16
País:Resto de Europa	111.8415457	7.8183	14.305	<2e-16
País:Bélgica	82.24963343	7.1023	11.581	<2e-16
País:Irlanda	57.30492426	9.3469	6.131	8.76E-10
País:Suiza	55.92887282	8.0448	6.952	3.61E-12
País:Francia	-0.952849053	5.7769	-0.165	0.868991
País:Reino Unido	-34.62213693	5.0322	-6.88	6.00E-12
País:Portugal	-77.36631374	11.2683	-6.866	6.63E-12
País:Italia	-121.7166981	6.5324	-18.633	<2e-16
mes:02	11.46124716	7.2695	1.577	0.114884
mes:03	25.26693936	7.1269	3.545	0.000392
mes:04	37.28832176	7.0311	5.303	1.14E-07
mes:05	49.65203915	6.951	7.143	9.16E-13
mes:06	78.30903393	7.0338	11.133	<2e-16
mes:07	151.5625705	6.61	22.929	<2e-16
mes:08	113.0740145	6.7014	16.873	<2e-16
mes:09	52.33552929	8.6385	6.058	1.38E-09
mes:10	22.46745458	7.1962	3.122	1.80E-03
mes:11	30.02920689	7.0804	4.241	2.22E-05
mes:12	33.93658951	7.1409	4.752	2.01E-06
Com. Aut. Andalucia	22.80569234	6.3856	3.571	0.000355
Com. Aut. Illes Balears	71.58490312	6.2605	11.434	<2e-16
Com. Aut. Galicia	-51.64013851	10.9135	-4.732	2.23E-06
Com. Aut. Comunitat Valenciana	-27.12316128	6.1677	-4.398	1.10E-05
Com. Aut. Cantabria	-34.35207071	15.1637	-2.265	0.023488
Com. Aut. Catalunha	-5.406620006	5.5292	-0.978	0.328157
Com. Aut. Castilla y Leon	-56.02347189	12.2549	-4.572	4.85E-06
Com. Aut. La Rioja	-100.8217133	29.2061	-3.452	5.56E-04
Com. Aut. Canarias	169.7889735	6.9021	24.599	<2e-16
Com. Aut. Castilla-La Mancha	-31.19470674	22.4222	-1.391	0.164155
Com. Aut. Pais Vasco	56.91480057	9.0717	6.274	3.53E-10
Com. Aut. Comunidad foral de Navarra	-35.0800416	20.9371	-1.675	0.093838
Com. Aut. Region de Murcia	5.84092462	11.8887	0.491	0.623216
Com. Aut. Extremadura	6.657943158	25.3133	0.263	0.792534
Com. Aut. Aragon	-23.2559943	14.5015	-1.604	0.108783
Com. Aut. Principado de Asturias	-83.14163047	15.818	-5.256	1.47E-07
Com. Aut. Ceuta	-199.2256368	193.3308	-1.03	0.302781
Com. Aut. Melilla	-104.9668203	160.8897	-0.652	0.514135
Vía Sal. Carretera	-380.7467449	4.9683	-76.635	<2e-16
Vía Sal. Tren	-136.6890257	11.9396	-11.448	<2e-16
Vía Sal. Puerto	-99.23280654	7.5676	-13.113	<2e-16
Aloj. Alquiler	408.7683797	5.0387	81.126	<2e-16
Aloj. Hoteles y similares	407.7740875	3.8574	105.712	<2e-16
Motivo: Negocios	86.1719956	5.9509	14.481	<2e-16
Motivo: Ocio/vacaciones	-40.47521729	4.2488	-9.526	<2e-16
Paq. Tur. (si)	-23.08010911	4.4038	-5.241	1.60E-07

Tabla 2: Correlación entre los residuos de cada par de métodos.

	LR	KNN	RF	SVM	DL
LR	1	0.756	0.859	0.824	0.789
KNN	0.756	1	0.809	0.768	0.675
RF	0.859	0.809	1	0.890	0.848
SVM	0.824	0.768	0.890	1	0.858
DL	0.789	0.675	0.848	0.858	1

Tabla 3: Importancia relativa de las variables para cada método. En cada caso se muestra el valor correspondiente al set de entrenamiento (Train) y el set de prueba (Test).

	Relevancia LR		Relevancia SVM		Relevancia DL		Relevancia RF		Relevancia KNN		Relevancia promedio	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Mes	0.08609	0.07709	-0.0486	-0.03674	-0.06384	-0.04552	-0.06856	-0.034	0.07031	0.10564	-0.00492	0.01329
Vía sal.	0.1137	0.1055	0.02424	0.0408	0.04696	0.05532	-0.03348	0.02074	0.01299	0.15024	0.03288	0.07452
País	0.41944	0.4052	0.39899	0.38941	0.46612	0.45935	0.38959	0.38152	0.55071	0.55116	0.44497	0.43733
Com. Aut.	0.08982	0.08126	0.02412	0.0376	0.04491	0.05567	-0.02016	0.01124	0.14947	0.18607	0.05763	0.07437
Tot. Pern.	0.79833	0.79124	0.852	0.84948	0.99624	1	0.65947	0.64608	0.58329	0.60881	0.77787	0.77912
Alojamiento	0.15532	0.14636	0.11609	0.13613	0.15803	0.17504	0.07777	0.11175	0.02988	0.15085	0.10742	0.14403
Motivo	0.08497	0.07575	-0.05009	-0.02734	-0.05849	-0.03536	-0.04396	0.0104	-0.02785	0.10511	-0.01908	0.02571
Paq. Tur.	0.08161	0.07274	-0.03258	-0.00159	-0.00561	0.02253	-0.09024	-0.01105	-0.06914	0.10151	-0.02319	0.03683

Apéndice B. Figuras

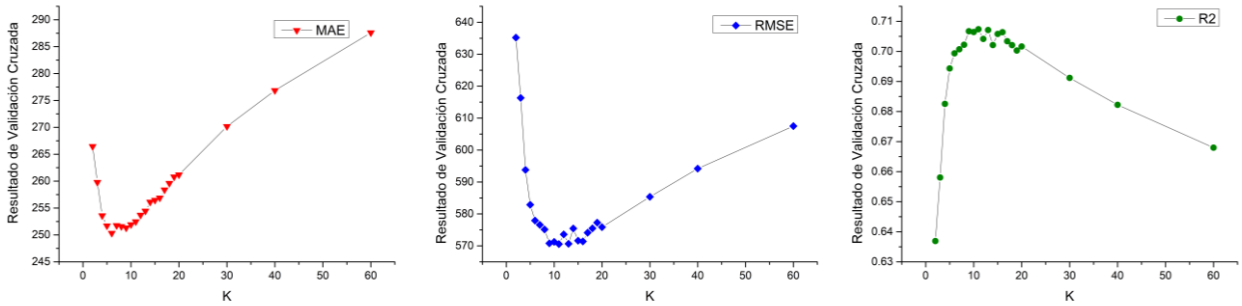

Figura 20: Dependencia de las diferentes medidas de la calidad del ajuste con el valor de K utilizado en el algoritmo KNN.

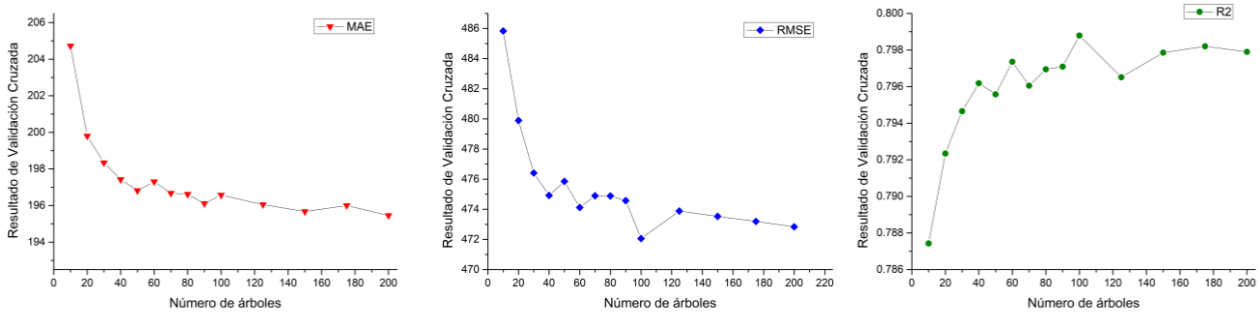

Figura 21: Dependencia de las diferentes medidas de la calidad del ajuste con el número de árboles utilizados en el algoritmo RF.